The Economics of the Cloud: Price Competition and Congestion

[Extended Abstract]

Jonatha Anselmi
Basque Center for Applied Mathematics

Danilo Ardagna
Dip. di Elettronica e Informazione, Politecnico di Milano

John C.S. Lui
Computer Science & Engineering, Chinese University of Hong Kong

Adam Wierman
Computing and Mathematical Sciences, California Inst. of Technology

Yunjian Xu
Computing and Mathematical Sciences, California Inst. of Technology

Zichao Yang
Computer Science & Engineering, Chinese University of Hong Kong

1. INTRODUCTION

The cloud computing marketplace has evolved into a highly complex economic system made up of a variety of services, which are typically classified into three categories:

(i) In *Infrastructure-as-a-Service (IaaS)*, cloud providers rent out the use of (physical or virtual) servers, storage, networks, etc. To deploy applications users must install and maintain operating systems, software, etc. Examples include Amazon EC2, Google Cloud, and Rackspace Cloud.

(ii) In *Provider-as-a-Service (PaaS)*, cloud providers deliver a computing platform on which users can develop, deploy and run their application. Examples include Google App Engine, Amazon Elastic MapReduce, and Microsoft Azure.

(iii) In *Software-as-a-Service (SaaS)*, cloud providers deliver a specific application (service) for users. There are a huge variety of SaaS solutions these days, such as email services, calendars, music services, etc. Examples include services such as Dropbox, Gmail and Google Docs.

This abstract aims to introduce and analyze a stylized model capturing the multi-tiered interaction between users and cloud providers in a manner that exposes the interplay of congestion, pricing, and performance issues. To accomplish this, we introduce a novel three-tier model for the cloud computing marketplace. This model, illustrated in Figure 1, considers the strategic interaction between users and SaaS providers (the first and second tiers), in addition to the strategic interaction between SaaS providers and either IaaS or PaaS providers (the second and third tiers). Of course, within each tier there is also competition among users, SaaS providers, and IaaS or PaaS providers, respectively. To the best of our knowledge, the results described here are the first to jointly consider the interactions and the equilibria arising from the full cloud computing stack (i.e., users, services, and infrastructures/platforms).

WPIN '13 Pittsburgh, PA USA

A key aspect captured by the model presented here is that the performance experienced by the users is modeled as a combination of congestion at *dedicated* resources, where congestion depends only on traffic from the SaaS, and *shared* resources, where congestion depends on the total traffic to the IaaS/PaaS.

Our analysis in the full version paper [5] highlights a number of important, novel qualitative insights. For example, our results highlight that SaaSs extract profits only as a result of dedicated latency; while IaaS/PaaS providers extract profits from both shared and dedicated latencies. However, the profit of IaaS/PaaS providers reduces significantly as competition grows, and converges to zero in the limit, while services remain profitable even when there are a continuum of services. This highlights that SaaS providers maintain market power over IaaS/PaaS providers even when services are highly competitive, and that one should not expect the cloud marketplace to support a large number of IaaS/PaaS providers. This observation is similar to the relationship of content providers to ISPs in the internet [1,2].

Another danger that our analysis highlights is that the market structure studied here can yield significant performance loss for users, as compared with optimal resource allocation. Specifically, the price competition among services and providers yields inefficient resource allocation, i.e., the price of anarchy can be arbitrarily large. This contrasts with the result in [4], where it is shown that the price of anarchy is one if each service has its own dedicated infrastucture. However, competition among PaaS/IaaS providers can result in significant improvements in user efficiency. In particular, as the number of providers (and thus competition) grows, in the limit we show that the price of anarchy cannot be higher than 2, when congestion costs are linear, and $k+1$ if congestion costs are polynomial with degree k.

2. MODEL OVERVIEW

This work proposes a model for studying the interaction among three parities in the cloud marketplace: users, service providers (services for short) and infrastructure providers (providers for short). Since, in practice, the number of providers is small (tens) and the numbers of services and users are huge, our model considers a finite number of providers P but treats services and users as infinitesimals in a non-

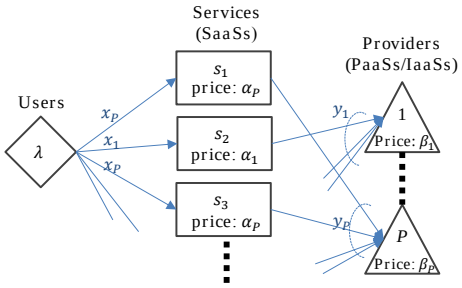

Figure 1: Overview of model structure and notation.

atomic model.

Providers: We consider P providers who sell resources to services, as done by Amazon EC2 and Google Cloud. The resources sold can represent virtual machines, in the case of an IaaS, or platforms provided for development, in the case of a PaaS. Each provider p charges a price β_p per unit of data flow for services that use its infrastructure. This charge-per-flow model is very common, e.g., it is used by Google App Engine. We let y_p denote the total flow of provider p and model the profit of provider p by

$$\text{Provider-Profit}(p) = \beta_p y_p, \tag{1}$$

Services: To study the aggregate behavior of a large number of cloud services, we consider a non-atomic model involving a continuum of infinitesimally small services, indexed by $s \in [0, 1]$. This modeling choice is appropriate since, in the real world, there are generally many more service providers than infrastructure providers. We assume that each service s chooses only one (infrastructure) provider, denoted by f_s, and that all services are homogeneous in the sense that the latency cost of users depends only on the provider chosen (not the service), which means that all services that choose the same provider are essentially identical from the standpoint of congestion. Further, since all services that choose the same provider are faced with the same profit-maximization problem, it is reasonable to assume that they charge the same price to their users. Therefore, we write the price charged by service s as α_{f_s}, which depends only on the provider it chooses. Since all users are cost-minimizing, it follows that all services that choose the same provider attract the same amount of data flow. We let x_p denote the flow of a service that chooses provider p. The profit of a service that chooses provider p as

$$\text{Service-Profit}(s) = (\alpha_p - \beta_p)x_p, \qquad \forall s : f_s = p.$$

Let g_p denote the fraction of services that choose provider p. We have $\sum_p g_p = 1$ and

$$\text{Provider-Profit}(p) = \beta_p x_p g_p.$$

Users: The customer base of cloud services is typically quite large, and so we use a nonatomic model in order to capture their aggregate behavior. We model the total user flow to the services as inelastic, and denote it by λ, i.e.,

$$\lambda = \sum_p g_p x_p = \sum_p y_p.$$

Latency in the cloud is determined by the combination of both the amount of flow at the service chosen, x_{f_s}, and the amount of flow using the provider chosen by the service y_{f_s}. Thus, we further break down the latency experienced into two types of congestion costs: 1) the *dedicated cost (latency)* from the service $\tilde{\ell}_{f_s}(x_{f_s})$ and 2) the *shared cost (latency)* from the provider $\hat{\ell}_{f_s}(y_{f_s})$. The dedicated cost represents congestion cost incurred at the service provider, e.g., due to the limited number of virtual machines held by the service. The shared cost represents the congestion at the infrastructure provider, e.g., the delay resulting from the network capacity shared by all services using the same infrastructure provider. Combining these latencies with the service price yields the "effective cost" that users seek to minimize. In particular, the effective cost of a user who chooses service s is

$$\text{User-Effective-Cost(s)} = \alpha_s + \tilde{\ell}_{f_s}(x_{f_s}) + \hat{\ell}_{f_s}(y_{f_s}). \tag{2}$$

3. EQUILIBRIUM CONCEPTS

The detailed definitions of the equilibria considered are beyond the scope of this abstract, but can be found in the full version of this paper [5]. In the following, we intuitively highlight the structure of the equilibria.

We assume that the users act at the fastest time scale, responding to fixed prices of the services and providers, and a fixed mapping of the services to the providers. The next fastest time scale we consider is pricing, with providers setting prices first and services responding optimally to them. Finally, the distribution of services to providers is modeled as the slowest time scale. This ordering is motivated by the behavior observed in practice: users move quickly between cloud services depending on price, service and provider prices also change quickly (hourly or faster), while the migration of services across providers happens infrequently.

In this context, we first fix the service distribution $\mathbf{g} = \{g_p\}_{p=1}^P$, and consider the equilibria of service and provider prices, $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\alpha_p, \beta_p\}_{p=1}^P$, according to a Stackelberg model where providers first set their prices and then services observe these prices and determine the prices they charge to end users. The user flow is then distributed according to a Wardrop equilibrium (cf. the latency cost defined in (2)).

Figure 2 shows an oligopolistic congestion game that mimics both user flow and (service and provider) profit resulting from a **price equilibrium** considered in the abstract. In this congestion game, each user has to go through two serial links to reach the "destination". An intermediate node represents a provider, and the p-th node attracts g_p fraction of services. The latency of each link is marked in Figure 2, which depends only on the total flow of the link.[1] Once a user chooses its service, its provider p is determined, and the user's cost is given by

$$\tilde{\ell}(x_p) + \hat{\ell}(y_p) + \gamma_p + \beta_p,$$

where $\gamma_p = \alpha_p - \beta_p$ can be regarded as the price charged by the light blue link the user chooses, and β_p is the price set by the dark blue link (the user's provider p). Since the congestion game depicted in Figure 2 has the same payoff structure as our model (with a fixed service distribution $\mathbf{g}$), the price equilibrium considered in the abstract essentially form a Stackelberg equilibrium of the congestion game where the P dark blue links choose their prices (simultaneously) at

[1]In contrast to a classical congestion game model, here the total flow of the p-th dark blue link (provider p) is $y_p = x_p g_p$.

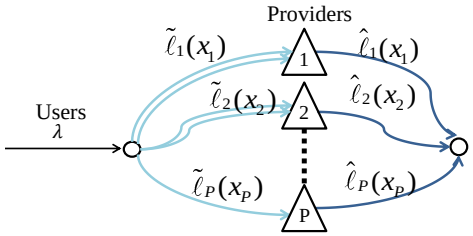

Figure 2: A congestion game that yields the same user flow (at a Wardrop equilibrium) as that resulting from a provider equilibrium of our model.

the first stage, and then at the second stage, all (non-atomic) light blue links set their prices.

The last component to incorporate into the definition is the mapping of services to providers, i.e., the distribution equilibrium, which fully characterizes the strategic interaction among the three market participants. We say a triple, $(\mathbf{g}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, is a **distribution equilibrium**, if: (i) service and provider prices $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ form a price equilibrium under $\mathbf{g}$; and (ii) no service has an incentive to change its provider because all providers yield services the same profit.

4. PROFITABILITY AND EFFICIENCY

A full analysis of the model can be found in the extended version of this paper [5]; however to give a flavor for the results, we consider two special cases in the following. These cases highlight insights that can be derived.

Profitability with Symmetric Providers

We consider a symmetric model where $\tilde{\ell}_p(x) = \tilde{a}x$ and $\hat{\ell}_p(y) = \hat{a}y$, for every p. There exists a symmetric distribution equilibrium such that $g_p = 1/P$, for every p. We have

$$\text{Provider-Profit(p)} = \left(\frac{2\tilde{a} + \hat{a}/P}{P-1}\right)\lambda^2, \qquad \forall p$$

and every service earns a profit of $\tilde{a}\lambda^2$.

These expressions for the provider and service profits are quite informative. In particular, they highlight that services extract profits only as a result of dedicated latency in this setting; while providers extract profits from both shared and dedicated latencies. However, competition among symmetric providers significantly reduces the profits providers can extract; as $P \to \infty$, provider profit goes to zero. In contrast, despite the fact that a continuum of services is considered, services still extract positive profit from the marketplace. This highlights that services maintain market power over providers even when services are highly competitive, and that one should not expect the cloud marketplace to support a large number of providers.

Price of Anarchy

The second question we study about the cloud marketplace is the effect of price competition in the cloud on the performance experienced by users. To study this question, we measure the "performance experienced by users" by the aggregate user latency resulting from a distribution equilibrium $(\mathbf{g}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. That is:

$$\ell(\mathbf{x}, \mathbf{g}) \triangleq \sum_p g_p x_p (\tilde{\ell}_p(x_p) + \hat{\ell}_p(g_p x_p)), \tag{3}$$

where $\mathbf{x} = (x_1, \ldots, x_P)$ is the Wardrop equilibrium resulting from the distribution equilibrium $\mathbf{g}$. We define the **price of anarchy (POA)** of a distribution equilibrium as the ratio of its resulting aggregate user latency to the minimum possible:

$$PoA \triangleq \frac{\ell(\mathbf{x}, \mathbf{g})}{\ell(\mathbf{x}^*, \mathbf{g}^*)}, \tag{4}$$

where $(\mathbf{x}^*, \mathbf{g}^*)$ is an optimal solution to the following optimization problem

$$\begin{aligned} \text{minimize}_{g\geq 0, x \geq 0} \quad & \ell(\mathbf{x}, \mathbf{g}) \\ \text{subject to:} \quad & \textstyle\sum_p g_p x_p = \lambda, \\ & \textstyle\sum_p g_p = 1, \end{aligned} \tag{5}$$

The next proposition provides an efficiency guarantee when all providers are nearly "symmetric".

PROPOSITION 1. *Assume $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ and $\hat{\ell}_p(y) = \hat{a}_p y^k$, for every p. Then,*

$$PoA \leq \frac{\tilde{a}_{\max} + \hat{a}_{\max}}{\tilde{a}_{\min} + \hat{a}_{\min}/P^k}, \tag{6}$$

where $\tilde{a}_{\min} = \min_p \tilde{a}_p$, $\hat{a}_{\min} = \min_p \hat{a}_p$, $\tilde{a}_{\max} = \max_p \tilde{a}_p$, and $\hat{a}_{\max} = \max_p \hat{a}_p$.

To explore the efficiency loss when the number of providers is large, we consider a "replica economy" scaling of providers where there are P types of providers and the number of providers of each type scales with n as n increases to infinity. In this context, as n increases to infinity, we show that there exists an ϵ-equilibrium (among all providers) with ϵ decreasing to zero. We show in [5] that the price of anarchy of a distribution equilibrium (on top of this ϵ price equilibrium) cannot exceed $k+1$.

This result highlights that the price of anarchy will be small in settings when there are a large number of providers. For example, the price of anarchy is bounded by 2 in the case of linear latencies. Interestingly, this is essentially the same price of anarchy as when no market structure exists, i.e., users directly choose providers based on congestion costs [3]. Since the price of anarchy of the two-tier model (users and SaaSs) converges to one in the limit as the number of services grows [4], this result reveals that the addition of providers into the marketplace "undoes" the efficiency created by competition among services.

References

[1] J. Musacchio, G. Schwartz, and J. Walrand, "A two-sided market analysis of provider investment incentives with an application to the net-neutrality issue", *Review of Network Economics*, vol. 8, no. 1, 2009.

[2] N. Economides and J. Tåg, "Network neutrality on the Internet: a two-sided market analysis", *Information Economics and Policy*, 2012.

[3] T. Roughgarden and E. Tardos, "How bad is selfish routing?", *Journal of the ACM*, vol. 49, 2002.

[4] J. Anselmi, U. Ayesta, and A. Wierman, "Competition yields efficiency in load balancing games", *Perform. Eval.*, vol. 68, no. 11, 2011.

[5] J. Anselmi, U. Ayesta, J. C.S. Liu, A. Wierman, Y. Xu and Z. Yang, "The economics of the cloud: price Competition and congestion", under submission, 2013.